# SENTIMENT ANALYSIS

by Ashutosh Kumar Singh and Abhishek Kamati

KIIT University | 2024

# ACKNOWLEDMENT

I would like to express my sincere gratitude to all those who have supported and guided me throughout the completion of the project "Intel Product Sentiment Analysis."

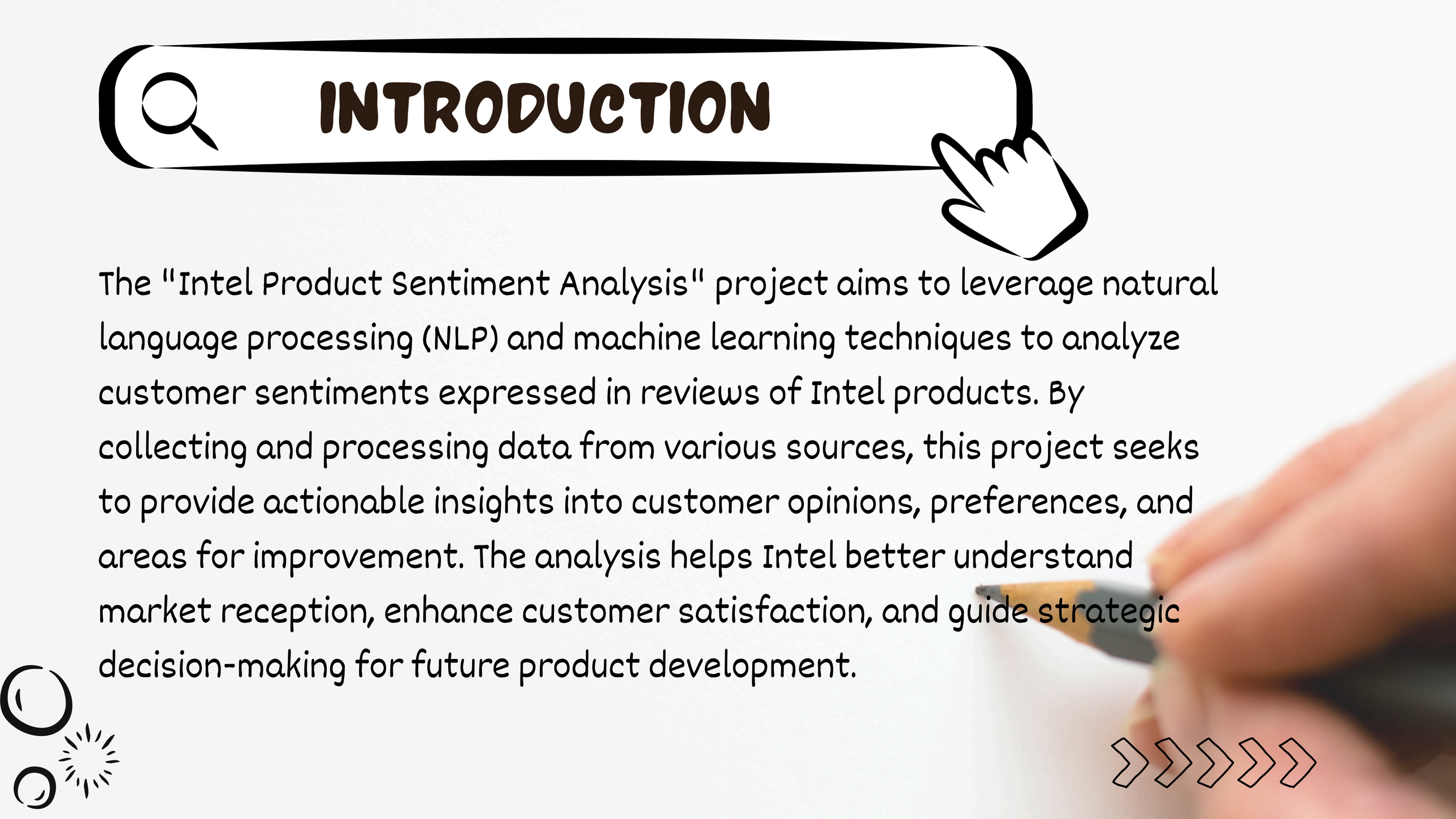
I extend my heartfelt thanks to our project manager, for providing clear instructions and unwavering support. Your leadership and guidance have been invaluable.

I am deeply grateful to my guide teacher, for your expert advice, encouragement, and insightful feedback. Your mentorship has been a cornerstone of this project's success.

Thank you all for your continuous support and belief in my abilities.

# INTRODUCTION

The "Intel Product Sentiment Analysis" project aims to leverage natural language processing (NLP) and machine learning techniques to analyze customer sentiments expressed in reviews of Intel products. By collecting and processing data from various sources, this project seeks to provide actionable insights into customer opinions, preferences, and areas for improvement. The analysis helps Intel better understand market reception, enhance customer satisfaction, and guide strategic decision-making for future product development.



# DATA COLLECTION

The first step towards building this project of ours was data collection. We generally relied on Amazon reviews for this data and collected this data via the scrapper developed by us, which you can see at this Scapper.py file in repo.
We collected data for each of the processors and thier generations as listed below.

# DATASETS

## i3

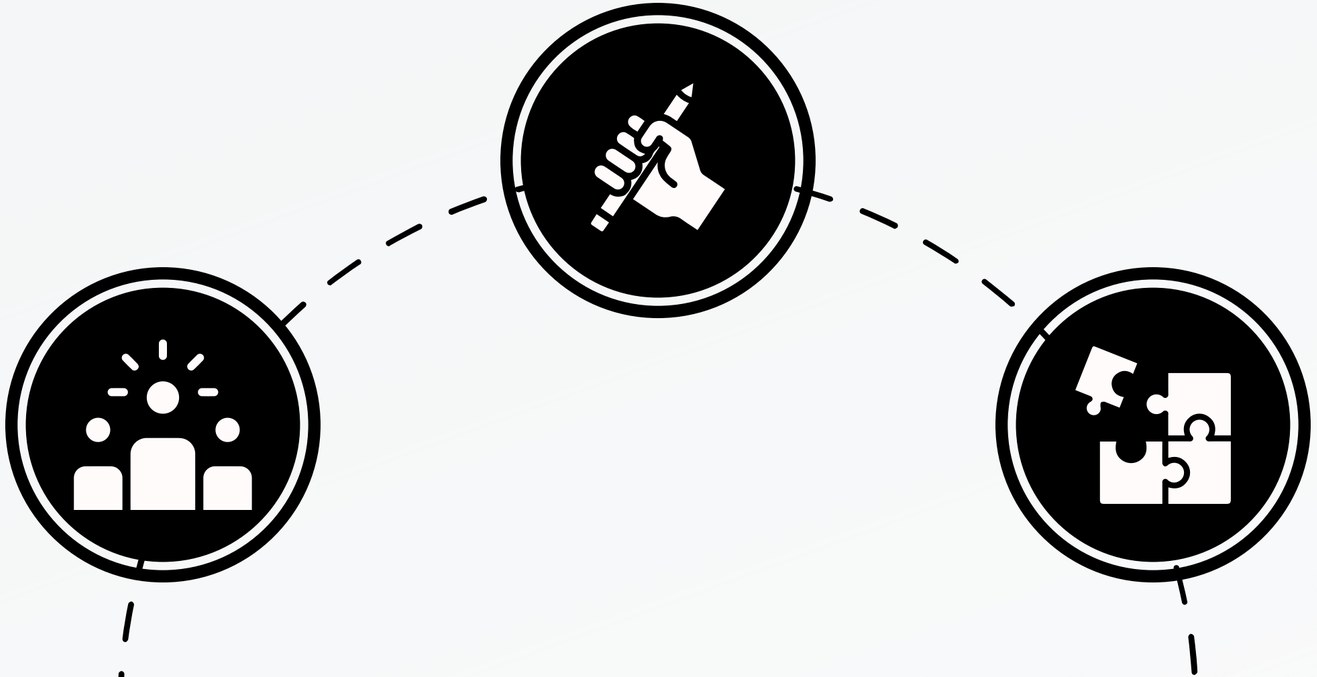
There are 4 datasets associated to i3:
1) i3 Gen 12
2)i3 Gen 13
3)i3 Gen 14

## i5

There are 4 datasets associated to i5:
1) i5 Gen 12
2)i5 Gen 13
3)i5 Gen 14

## i7

There are 4 datasets associated to i7:
1) i7 Gen 12
2)i7 Gen 13
3)i7 Gen 14



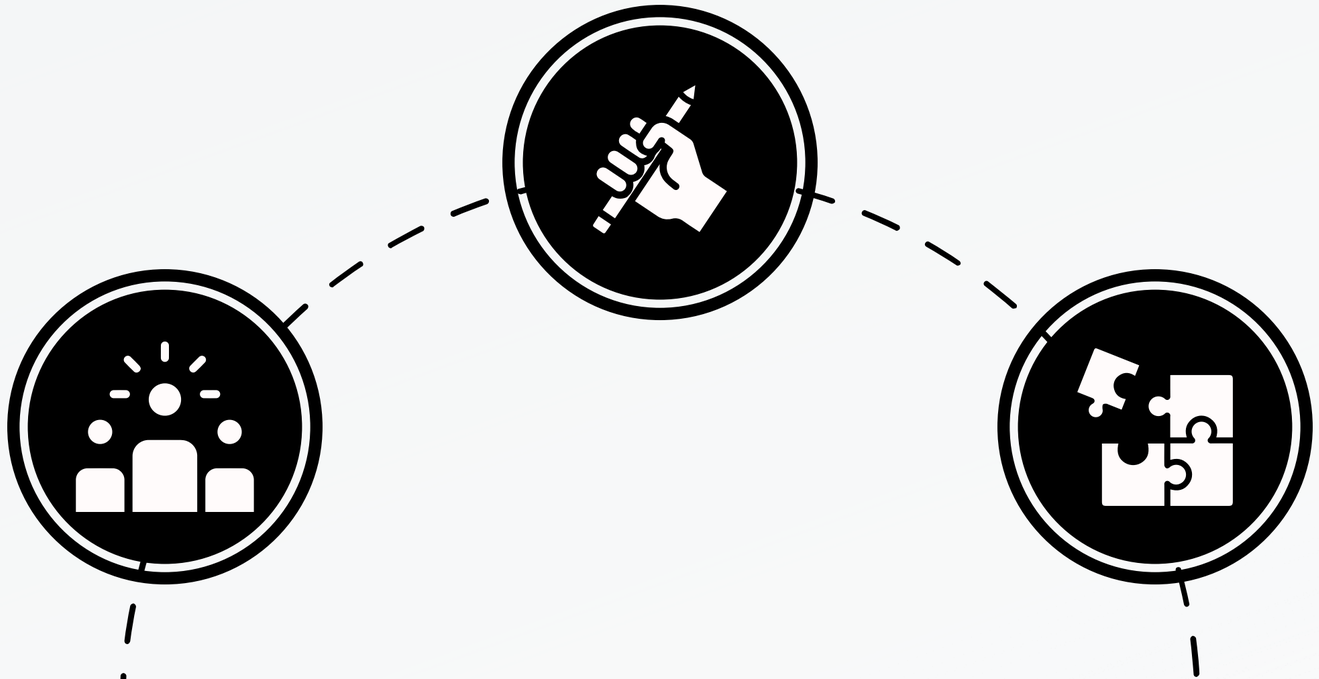
# DATASETS

## i9

There are 4 datasets associated to i9:
1) i9 Gen 12
2)i9 Gen 13
3)i9 Gen 14

## All Processor Dataset

This dataset is the combination of each of the datasets mentioned prior



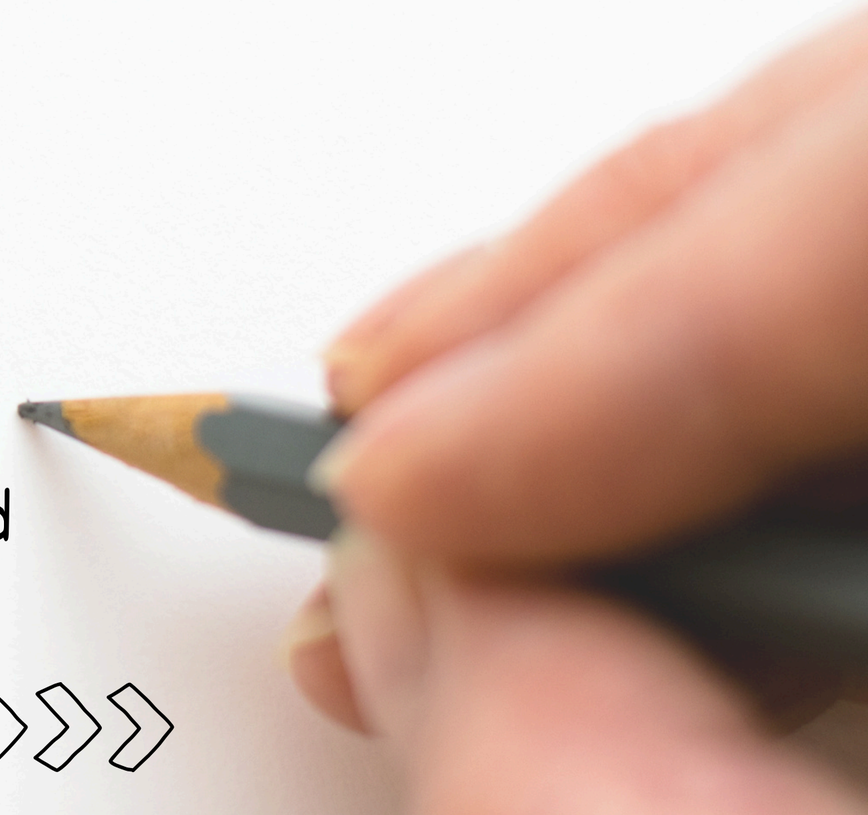
All this data is present in the data folder.

# DATA PREPROCESSING

The next step of this project was to process the data we had collected. In the upcoming text we will explain the steps involved in the proprocessing of data.

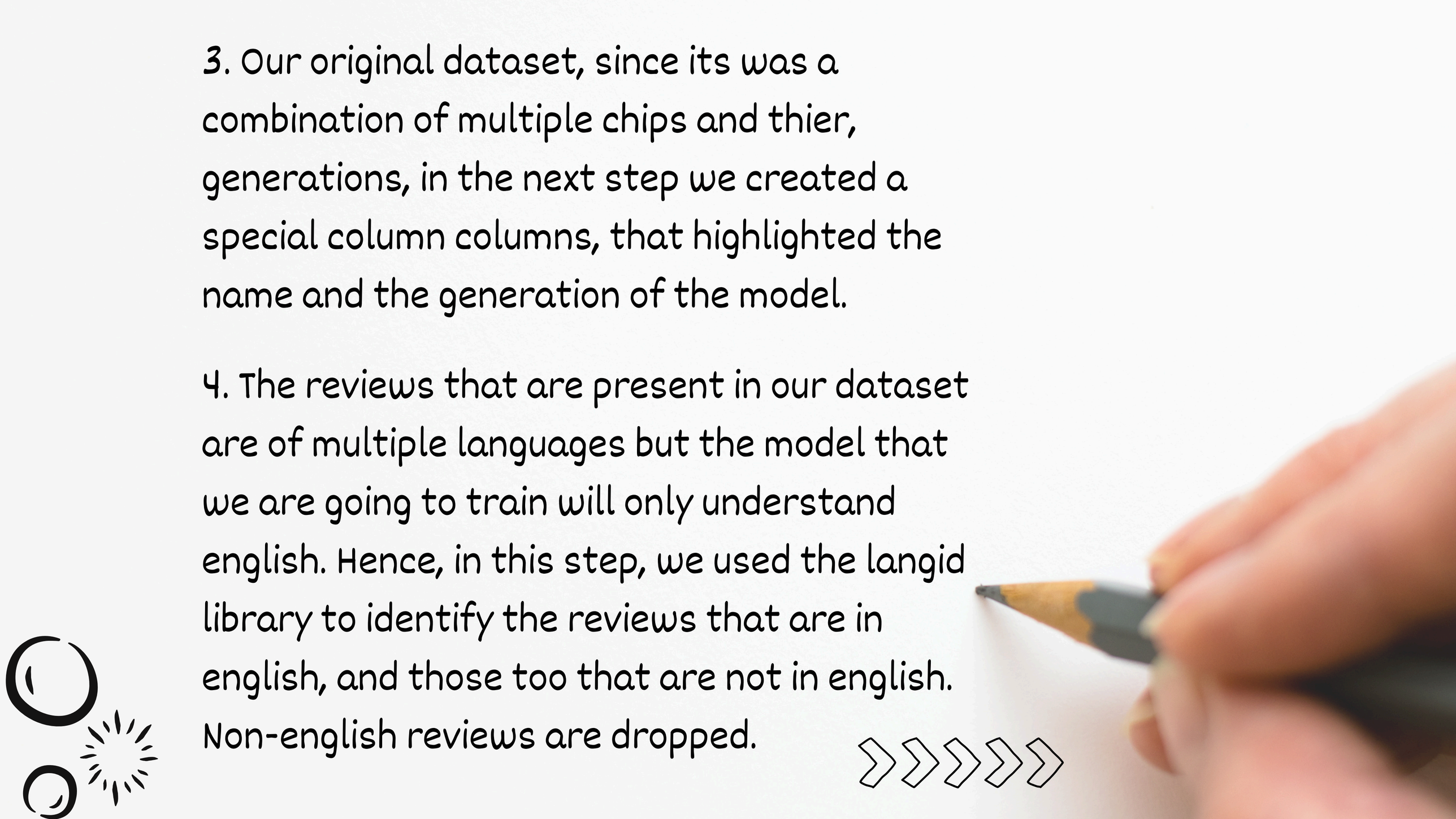
1. The first step was to remove the unwanted columns our dataset had.

2. The next step was to drop all the null valued rows from our dataset had.



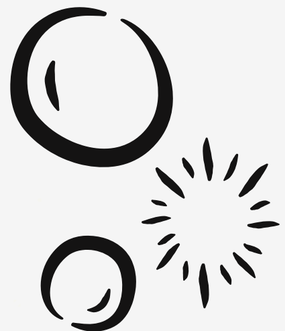
3. Our original dataset, since its was a combination of multiple chips and thier, generations, in the next step we created a special column columns, that highlighted the name and the generation of the model.

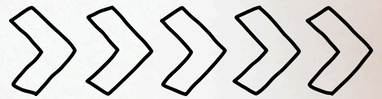
4. The reviews that are present in our dataset are of multiple languages but the model that we are going to train will only understand english. Hence, in this step, we used the langid library to identify the reviews that are in english, and those too that are not in english. Non-english reviews are dropped.

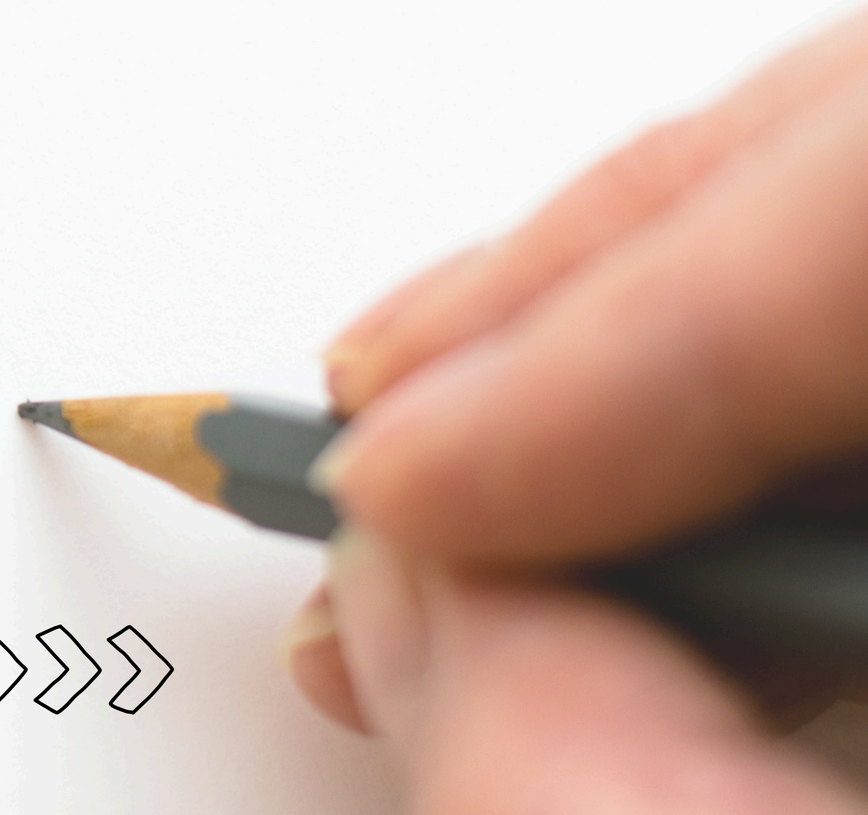


5. The next step of preprocessing was to then restore the index of the filtered dataset.

6. You can find how we performed data pre-processing via accessing notebook available in the Data_Preprocessing_Models folder.





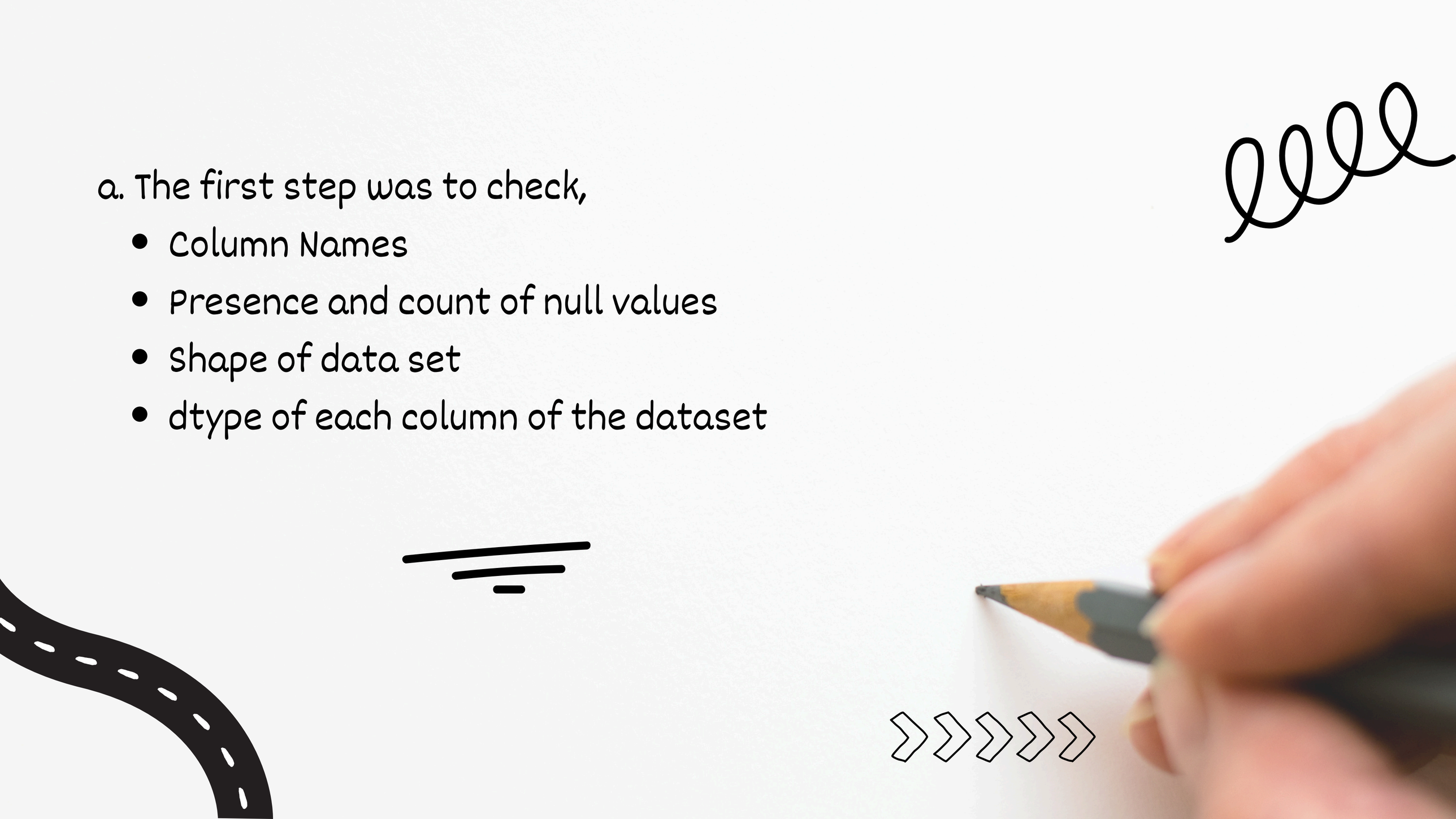


# SENTIMENT ANALYSIS METHODOLOGY AND RESULTS

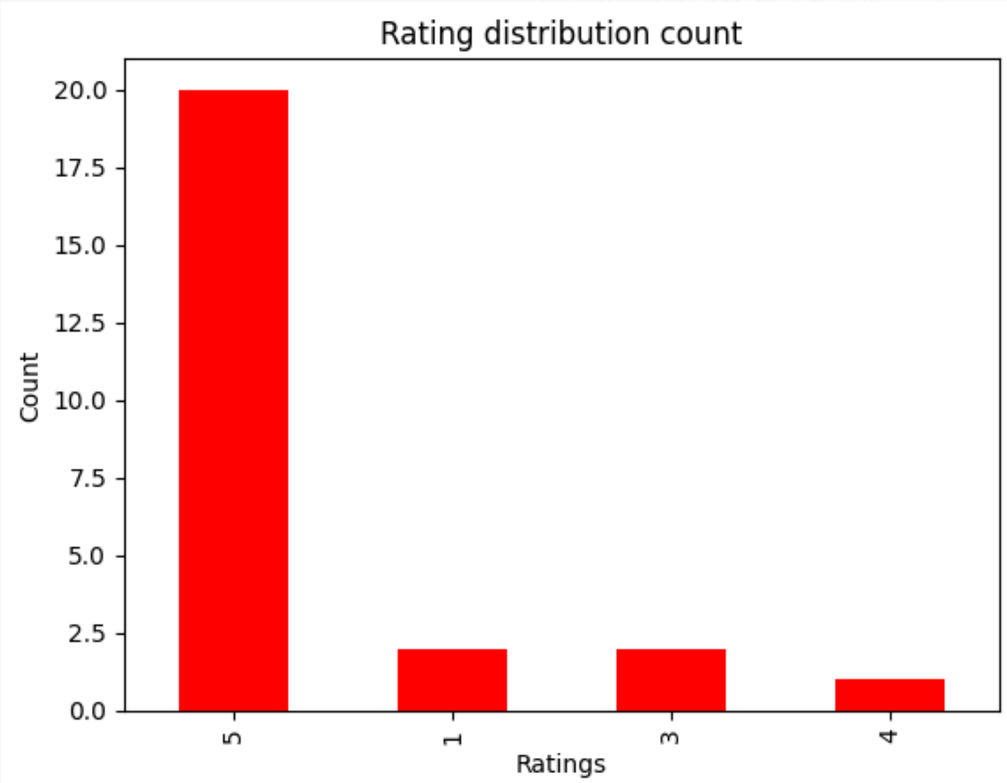
After data collection and preprocessing, the next step was to understand the data. For that we perform the EDA. We have already mentioned, that there are 13 different datasets, 4 of each of i3, i5 ,i7 and i9 and their different generations, and then a combined databases combining them all. We have performed EDA on each of these datasets separately. The following enlist the activities we did while performing the EDA.

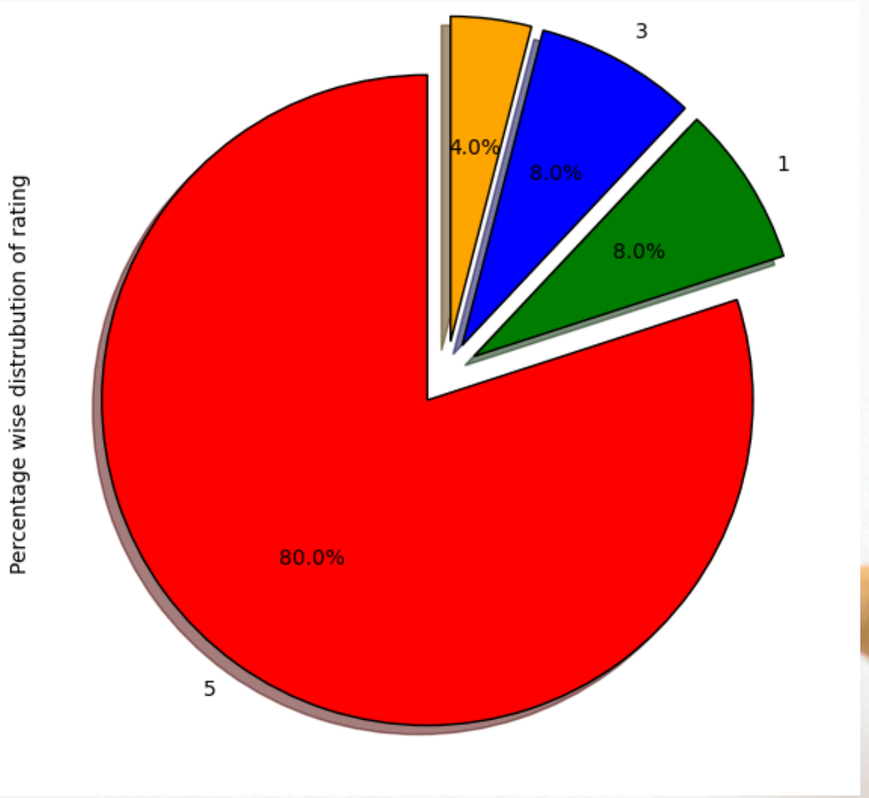
a. The first step was to check,

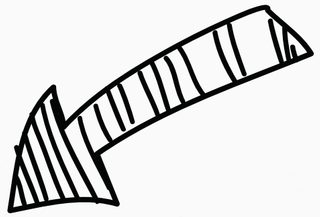
- Column Names
- Presence and count of null values
- Shape of data set
- dtype of each column of the dataset



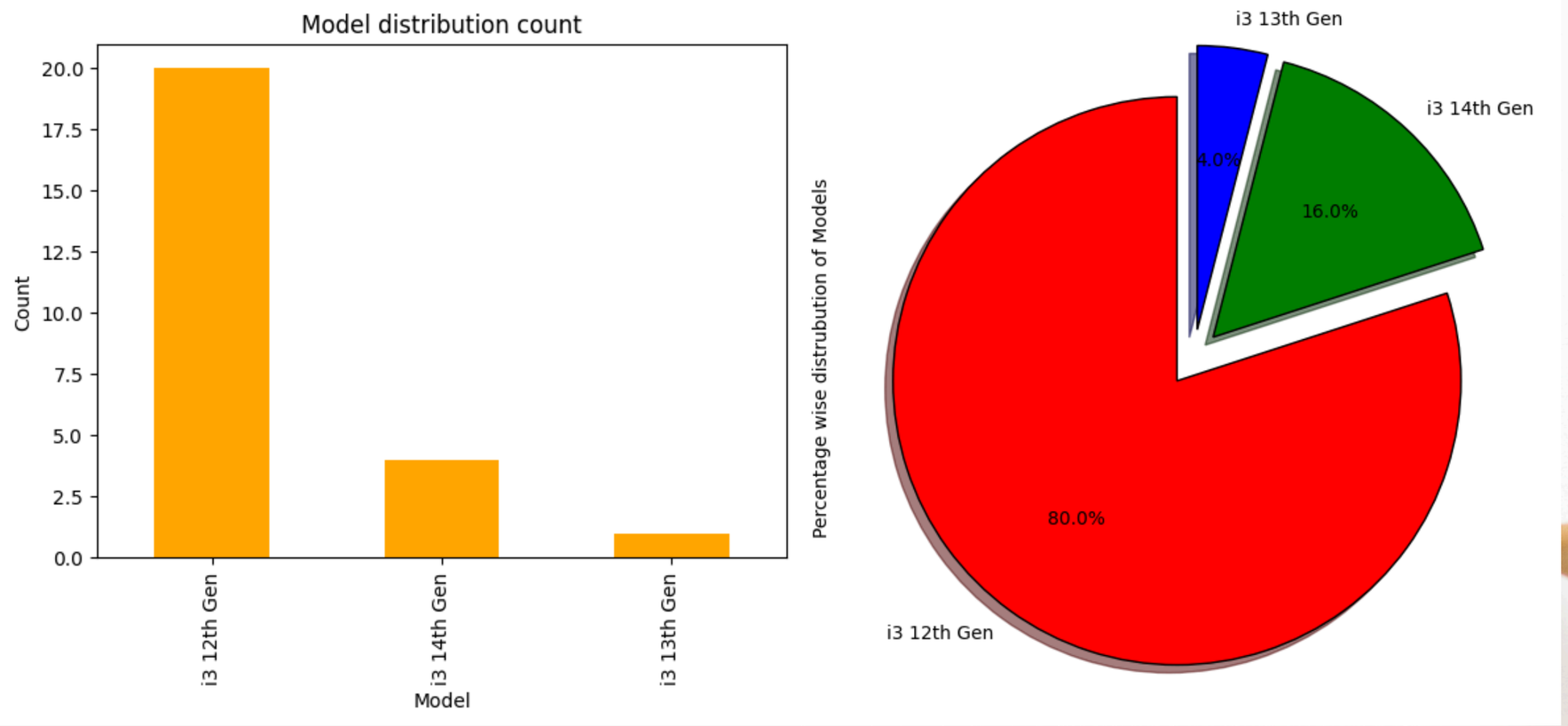
b. Next step was to analyze the count of each of the reviews belonging to the different ratings, that 1,2,3 ,4, and 5 and visualize this graphically.



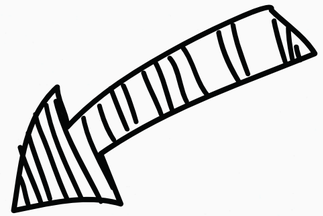



c. Next we have analyzed the count of reviews belong to each of the generations of the processors, that is, 12th gen, 13th gen and 14th gen.

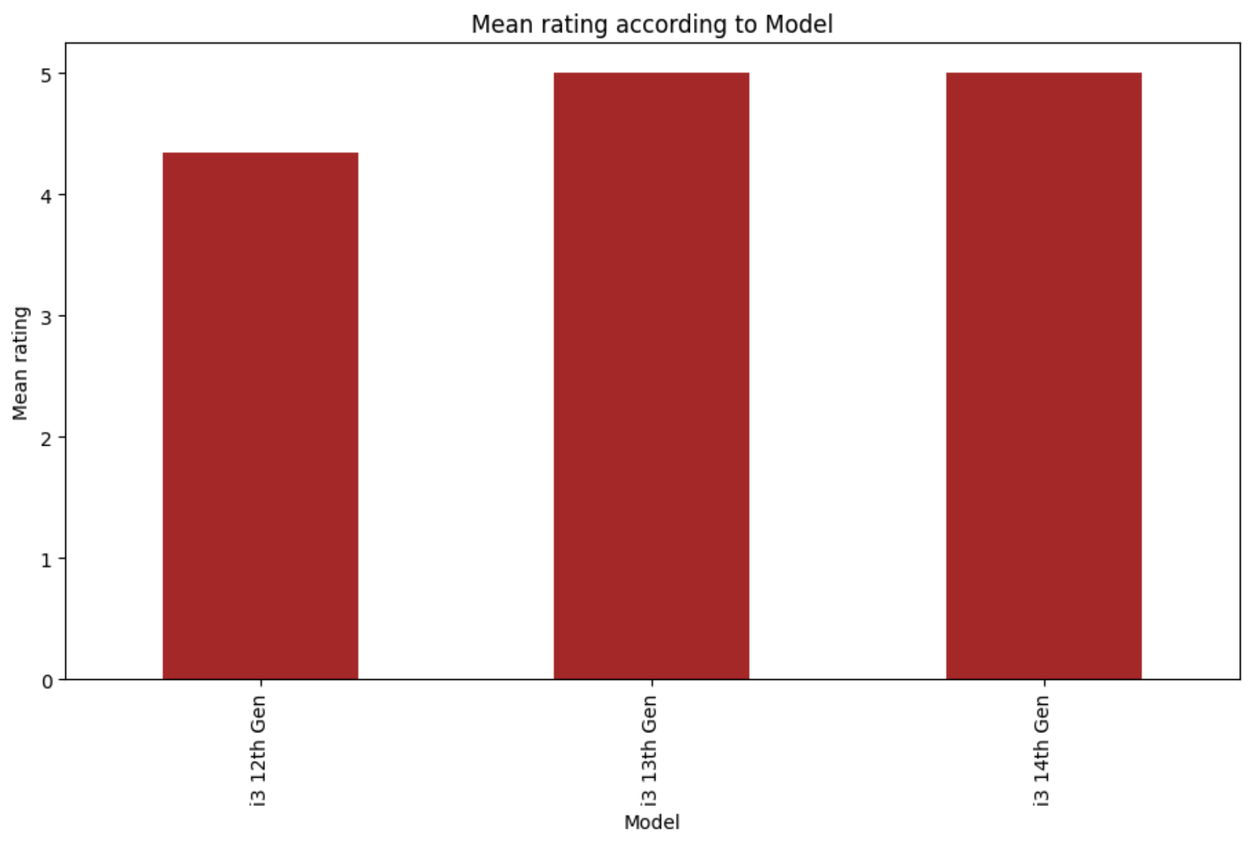


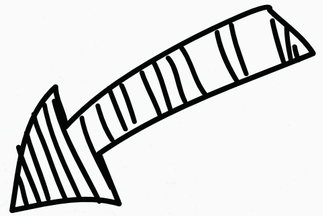




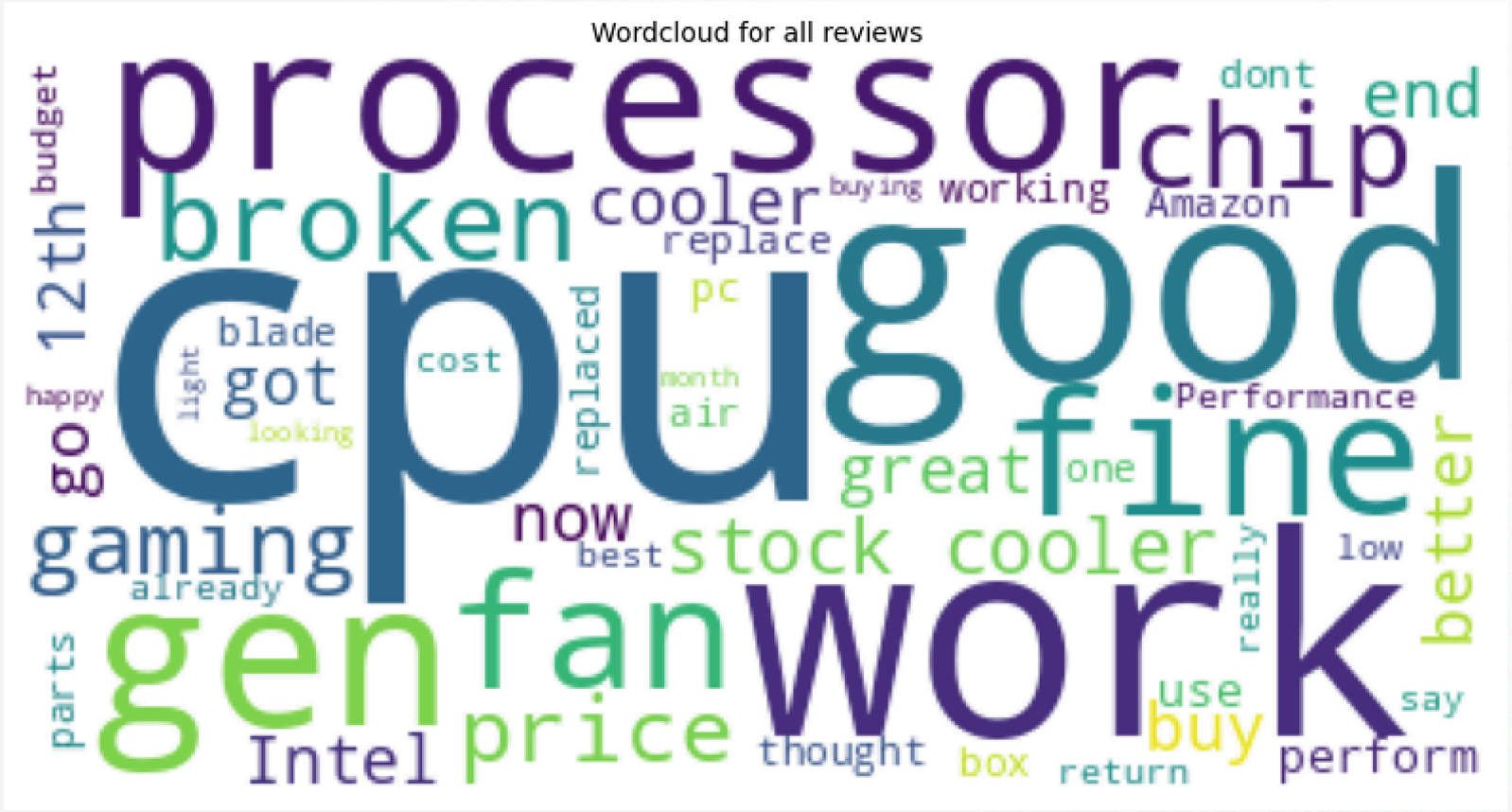


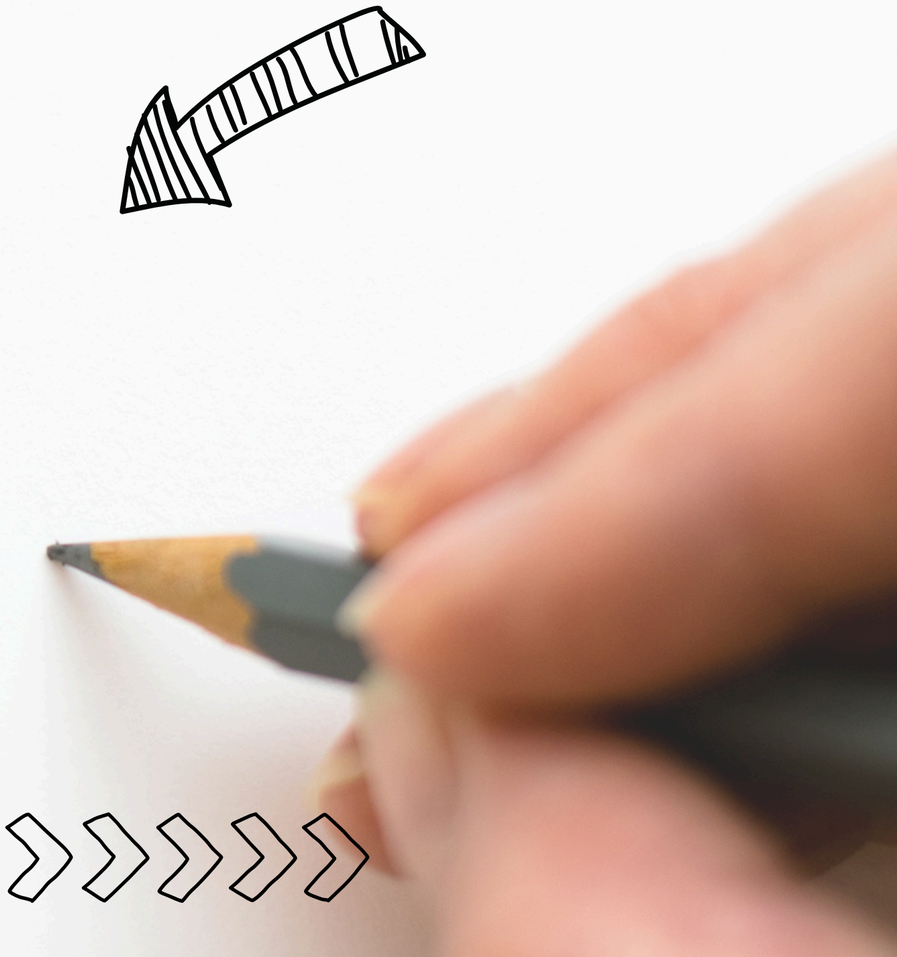
d. Next we have analyzed the  of average rating of each of generations of the processor and visualized it.



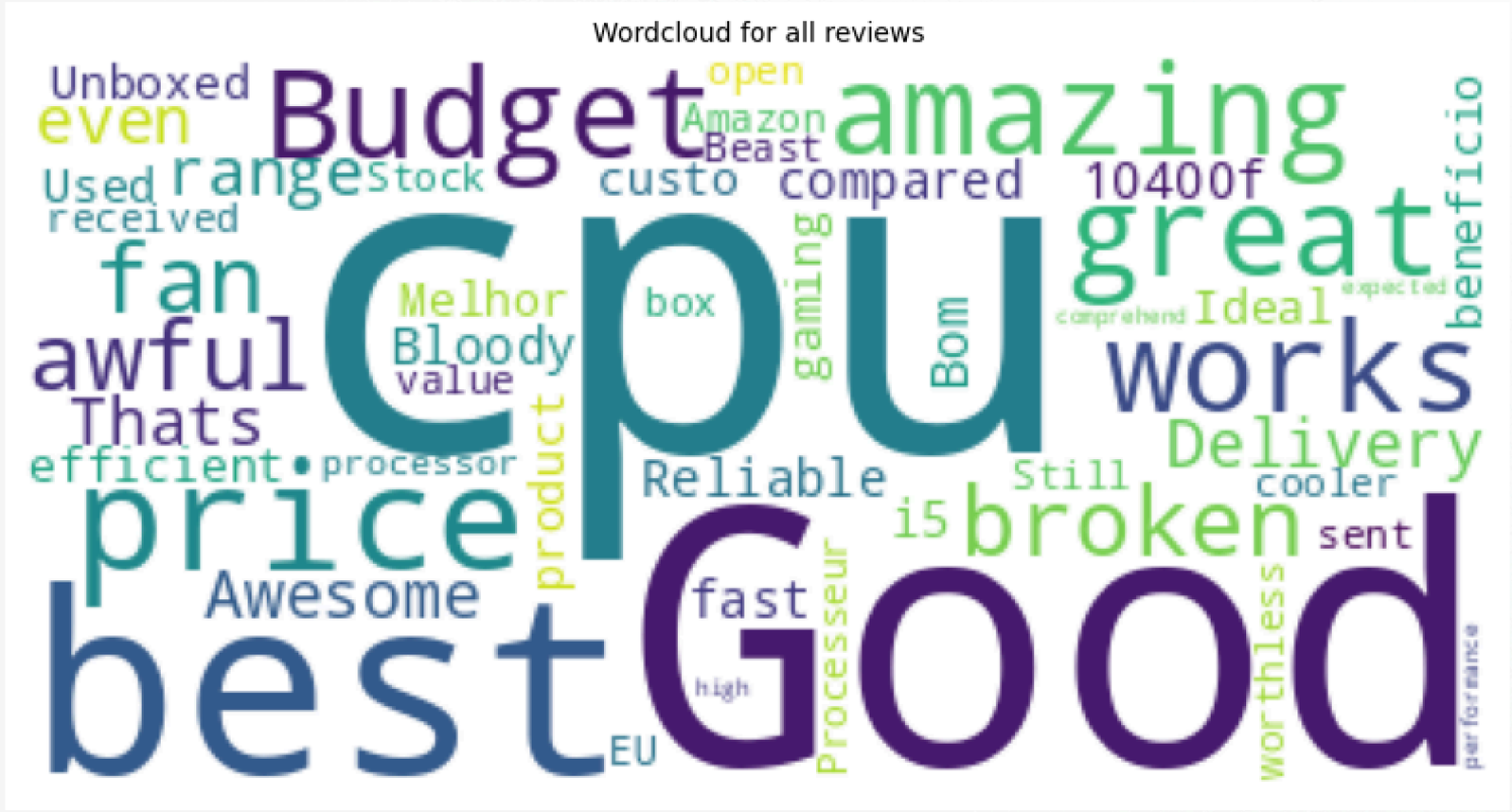



e. Next we have analyzed the reviews title column and generated the most used words by the users in review title of each of the reviews for the processors and their different generations

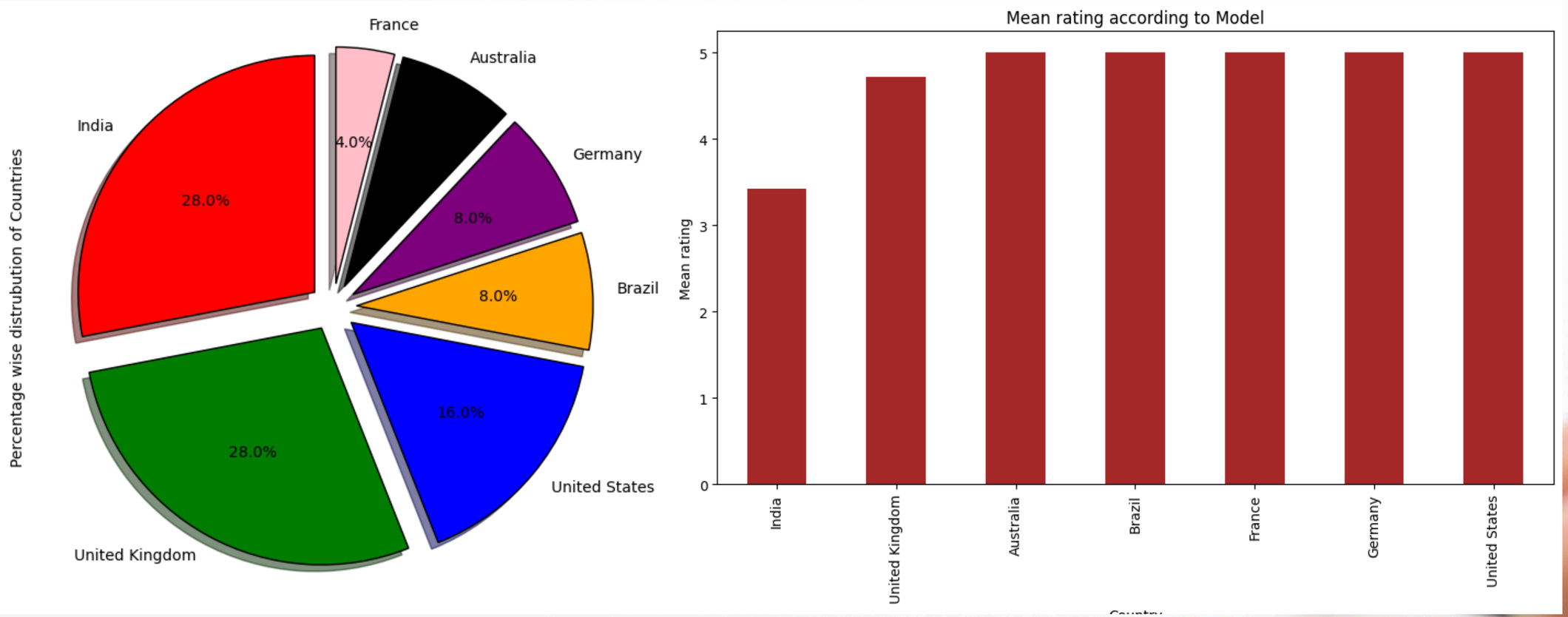


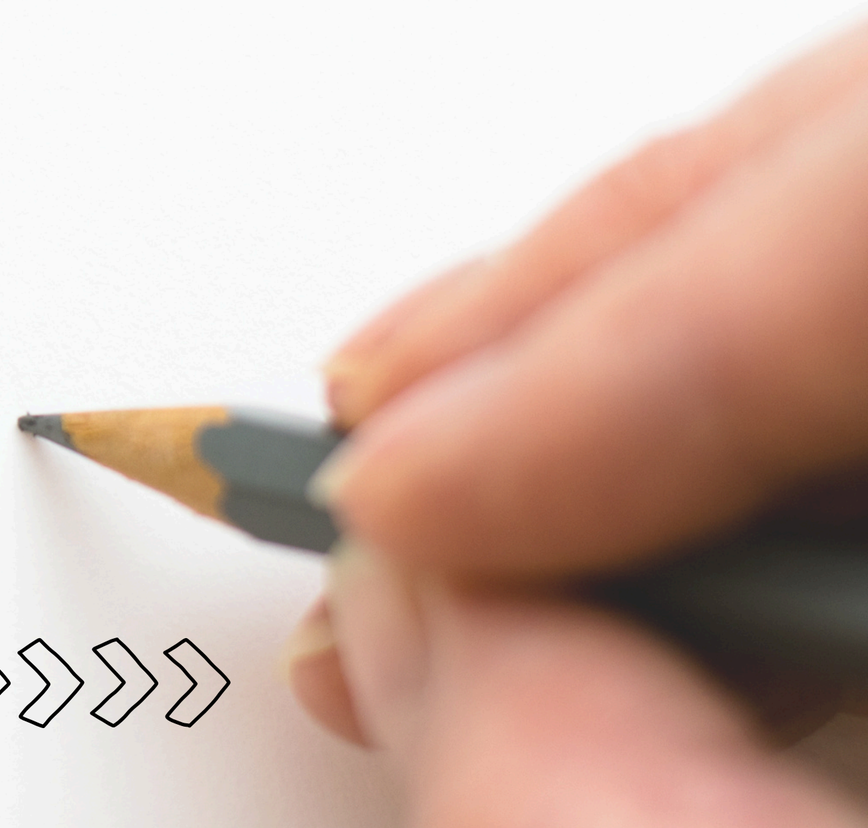
f. Next we have analyzed the reviews column and generated the most used words by the users in review of each of the processors and their different generations.



g. Next we have analysed the country column and analysed review count per country and average ratings per country.



After collections, preprocessing and EDA, the next step was to make a model that can analyze sentiments from the given text. Explaination following this slide, refers to steps taken to make a model that performs reasonably on input data.



The first step in making model that can analyse sentiments on normal reviews was to collect a new random review/text dataset, test a variety of models on that data compare their thier performances and choose the best one.

For that we downloaded Twitter Tweet Dataset. This dataset was then pre-processed and then we trained models on this dataset.

The models trained included the following:

- XGBoost
- Decision Tree
- Random Forest
- LSTM
- VADER
- RoBERTa

The test and train data accuracies of each of these models are as follows,

XGBoost

training accuracy: 0.6683302141817425

testing accuracy: 0.6343238326258338

Random Forest

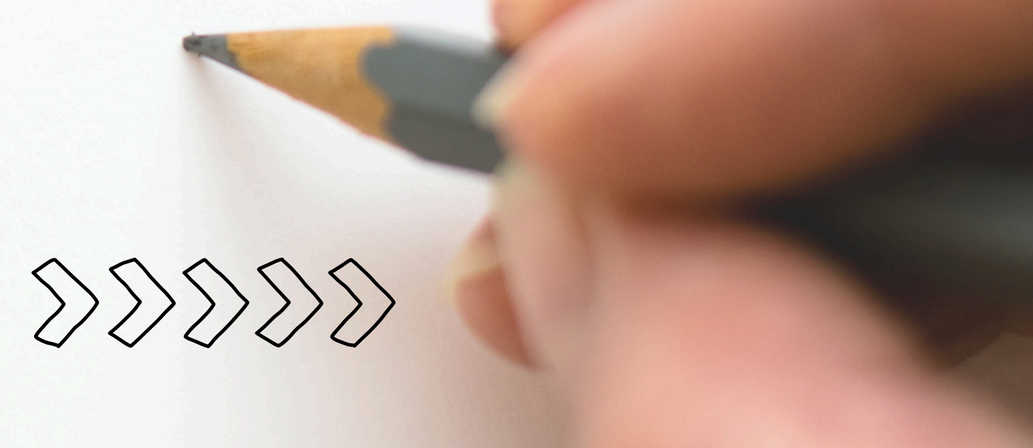
training accuracy: 0.9971407777084633

testing accuracy: 0.6357792601576713

Decision Tree

training accuracy: 0.6683302141817425

testing accuracy: 0.6343238326258338



The test and train data accuracies of each of these models are as follows,

LSTMs

training accuracy: 0.3245

testing accuracy: 0.3561

VADER

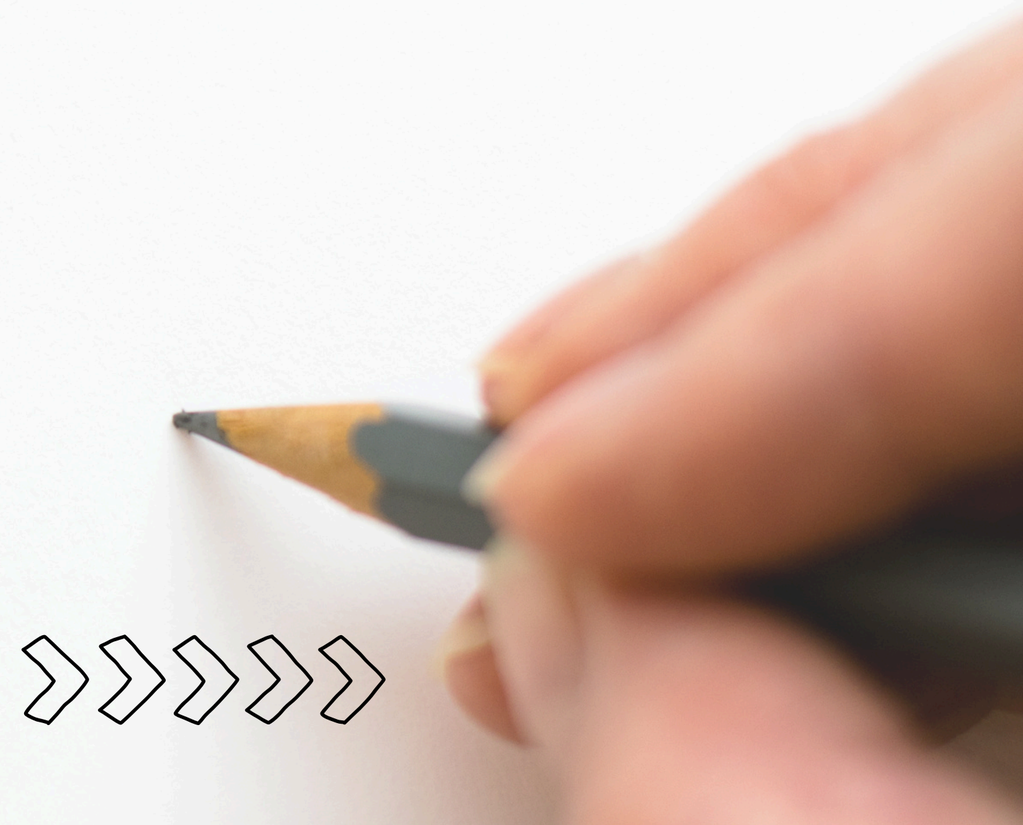
training accuracy: 0.65

testing accuracy: 0.63

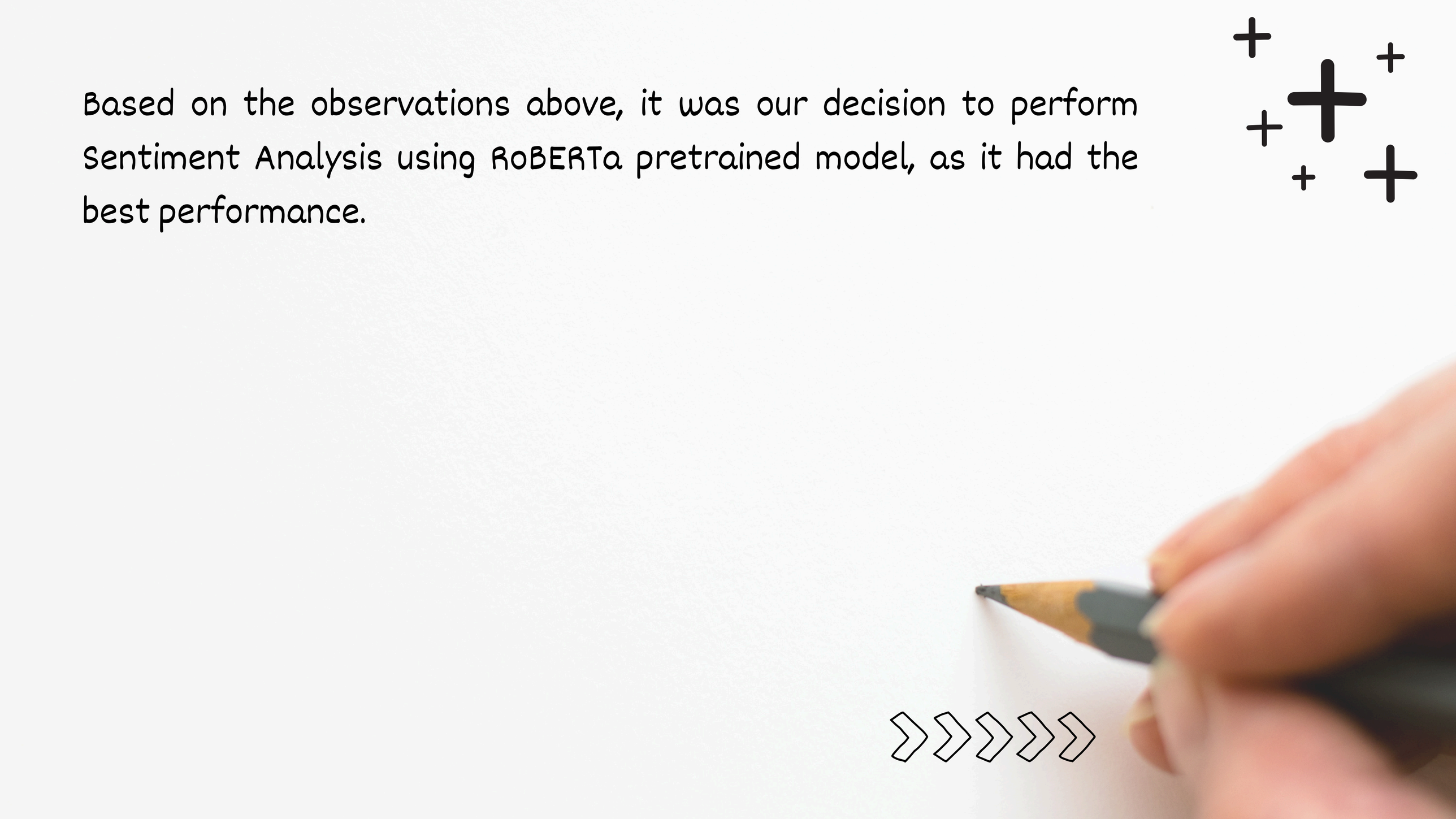
RoBERTa

training accuracy: 0.75

testing accuracy: 0.725



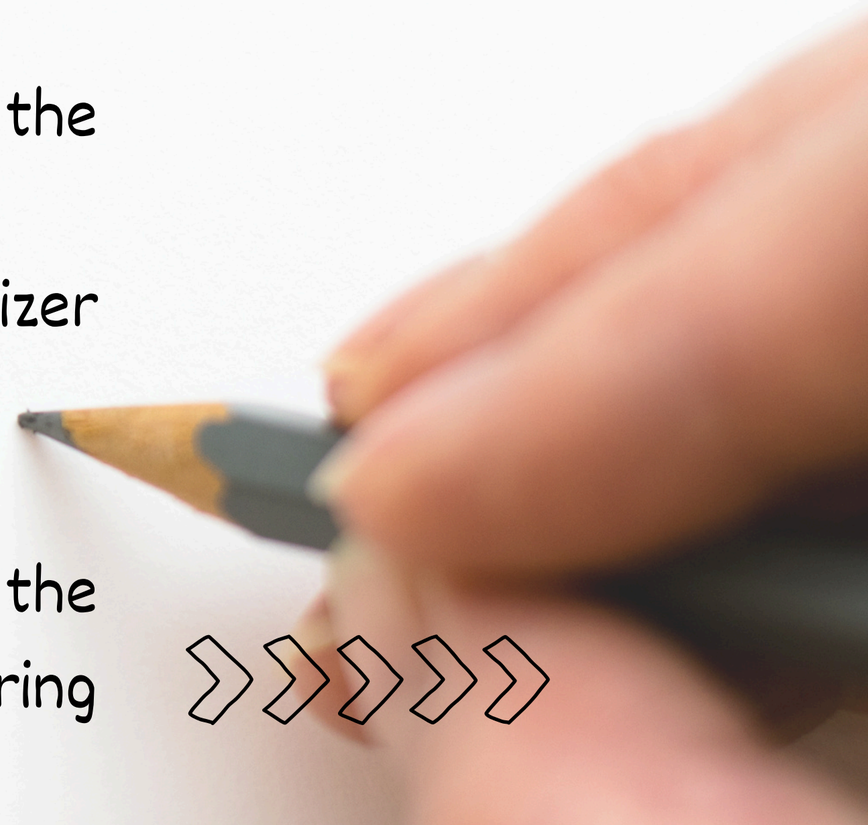
Based on the observations above, it was our decision to perform Sentiment Analysis using RoBERTa pretrained model, as it had the best performance.



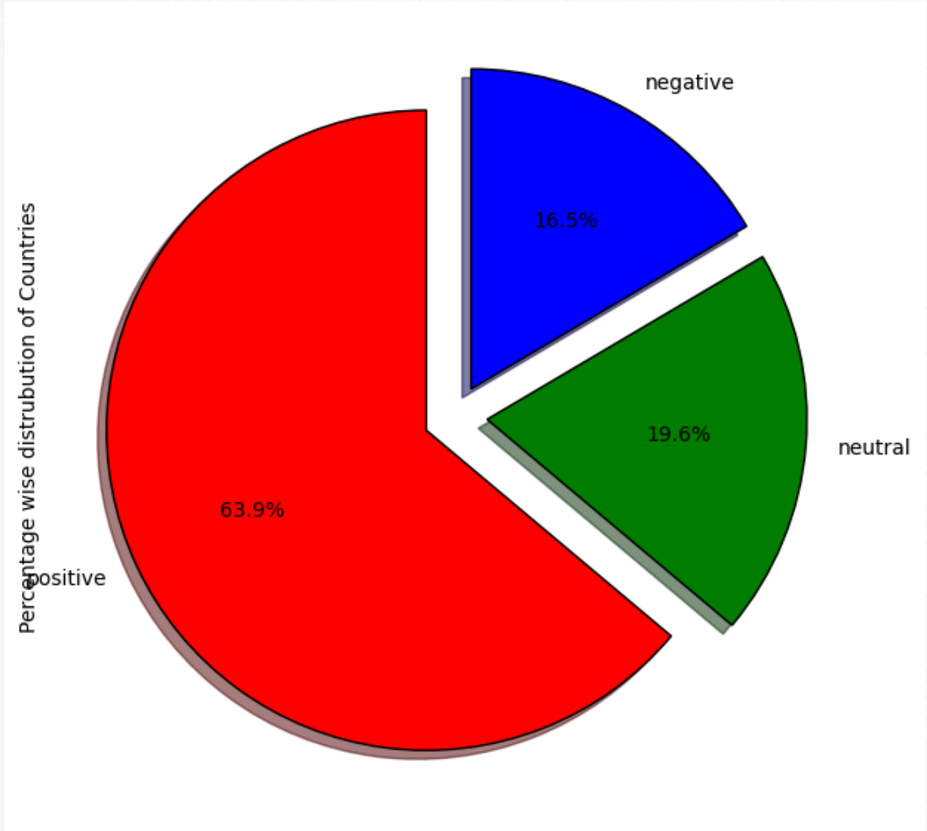
Now that our model to analyse sentiments is built and selected we will check for sentiments attached to each of the reviews on the intel processors. Now before we feed the reviews into the model, we need to pre-process the reviews. The steps taken are as follows,

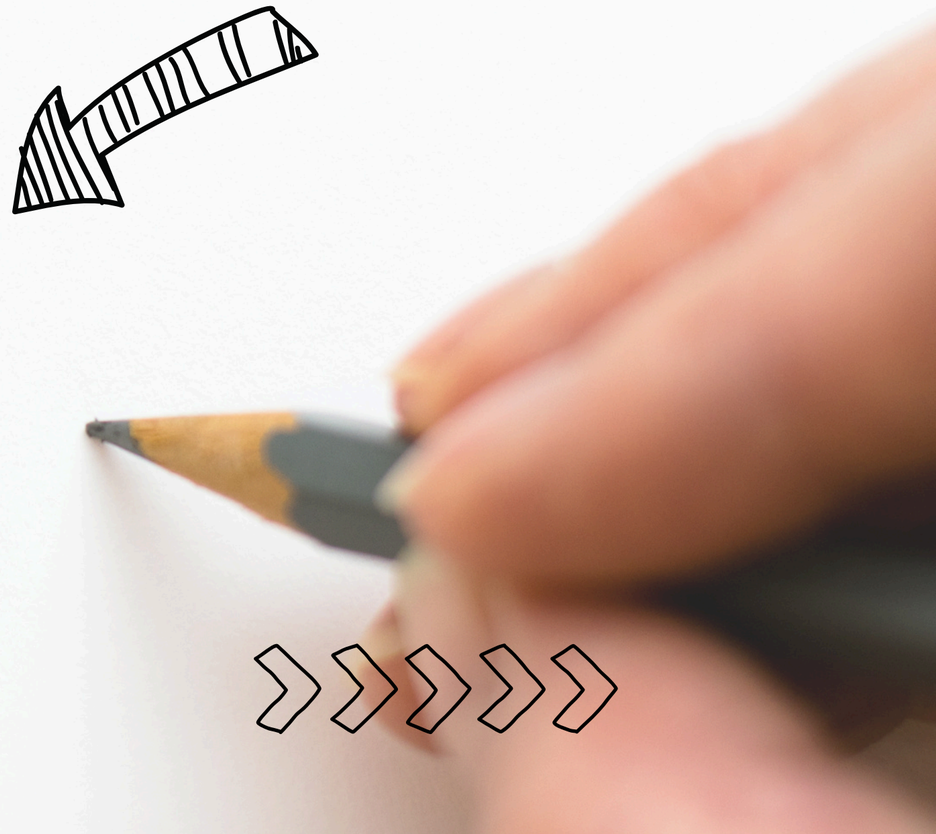
- Remove every character that isn't between a-z and A-Z.
- Remove stopwords.
- We perform stemming lemmatisation and extract the base words.
- We then perform vectorisation using CountVectorizer and bag of words approach.
- We then scale the input using min-max scaler.

And now our data is ready to be fed into model. Note the similar pre-processing technique was applied on data during model pre-processing.



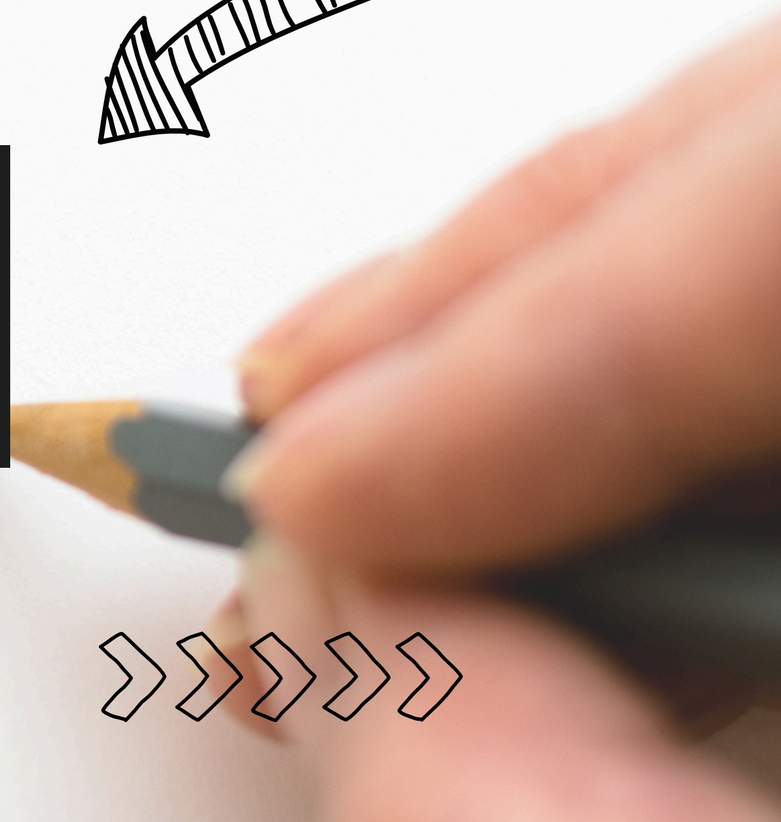
We now perform analysis of the sentiment results obtained after passing the reviews through Sentiment Analysis Model. We have performed it seperate on each generation of each processor, and then all of them combined. Below we see an example, of percentages, of positive, negative and neutral reviews of all processor and generation dataset.



Further we try to derive potential issues with each processor and its generation encountered by the users. For this we use the GenAI's API. We select the Negative and Neutral reviews from our user. We do this separately each processor and its generations. One of the instance can be seen below. Further can be seen in the file EDA_And_Modeling.

```
* **Excessive Heat and Thermal Throttling:**
    * Consistently high temperatures even with liquid cooling (100°C under load)
    * Spiking temperatures at idle (25°C to 70°C)
    * Thermal throttling during gaming, leading to stutters
    * Requires high-end cooling solutions (360mm AIOs, custom water loops)
    * Users recommend using CPU brackets to improve temps
    * Users report needing to manually configure BIOS settings to limit power draw and prevent instability
```



You can find all this, whether it be EDA, Sentiment Analysis or Model Selection in the files, Sentiment_Model_Selection and EDA_and_Modeling

# CONCLUSION

The Intel Product Sentiment Analysis project successfully leveraged natural language processing and machine learning techniques to evaluate customer sentiments towards Intel products. By analyzing customer reviews, we were able to categorize sentiments into positive, neutral, and negative classes, providing valuable insights into consumer perceptions. The implementation of models such as LSTM and RoBERTa, along with tools like VADER, allowed for robust sentiment classification. The project highlights the effectiveness of automated sentiment analysis in understanding customer feedback, which can aid Intel in enhancing product quality and customer satisfaction.


by Ashutosh Kumar Singh and
Abhishek Kamati